

# NCBI Resources: from Sequence to Function

Medha Bhagwat, NCBI

Current Topics in Genome Analysis  
September 9, 2003



## Outline

About NCBI  
NCBI Databases and Tools  
Example



# National Center for Biotechnology Information

<http://www.ncbi.nlm.nih.gov/>

Created as a part of NLM in 1988

- To establish public databases

U.S. National DNA Sequence Database

- To perform research in computational biology
- To develop software tools for sequence analysis
- To disseminate biomedical information



QUICK LINKS TABLE: ALPHABETICAL LIST WITH DIRECT LINKS TO RESOURCES			
<a href="#">About NCBI</a>	<a href="#">e-PCR</a>	<a href="#">MMDB</a>	<a href="#">Seminars</a>
<a href="#">ASN.1</a>	<a href="#">Entrez</a>	<a href="#">Model Maker <small>NEW</small></a>	<a href="#">Sequin</a>
<a href="#">BankIt</a>	<a href="#">ETP</a>	<a href="#">Mutation Databases (external)</a>	<a href="#">Site Search <small>NEW</small></a>
<a href="#">BLAST</a>	<a href="#">GenBank</a>	<a href="#">NCBI Handbook <small>NEW</small></a>	<a href="#">SKYCOGH</a>
<a href="#">Books</a>	<a href="#">GenBank sample record</a>	<a href="#">NCBI Home</a>	<a href="#">Software Engineering</a>
<a href="#">CDART</a>	<a href="#">Genes and Disease</a>	<a href="#">NCBI News</a>	<a href="#">Spidey</a>
<a href="#">CDD</a>	<a href="#">Genomic Biology</a>	<a href="#">Nucleotide Sequences (Entrez)</a>	<a href="#">Structures</a>
<a href="#">CGAP</a>	<a href="#">GEO (Expression)</a>	<a href="#">OMIM</a>	<a href="#">Submit Data</a>
<a href="#">Clones</a>	<a href="#">Glossary</a>	<a href="#">ORF Finder</a>	<a href="#">Taxonomy</a>
<a href="#">Cn3D</a>	<a href="#">HTGs</a>	<a href="#">Plant Genomes</a>	<a href="#">Tools</a>
<a href="#">Coffee Break</a>	<a href="#">HomoloGene</a>	<a href="#">Protein Sequences (Entrez)</a>	<a href="#">TPA <small>NEW</small></a>
<a href="#">COGs</a>	<a href="#">Human Genome Resources</a>	<a href="#">PROW</a>	<a href="#">Trace Archive</a>
<a href="#">Computational Biology Branch</a>	<a href="#">Human-Mouse Homology Maps</a>	<a href="#">PubMed</a>	<a href="#">UniGene</a>
<a href="#">dbEST</a>	<a href="#">LinkOut</a>	<a href="#">PubMed Central</a>	<a href="#">UniSTS</a>
<a href="#">dbGSS</a>	<a href="#">LocusLink</a>	<a href="#">RefSeq</a>	<a href="#">VAST</a>
<a href="#">dbSNP</a>	<a href="#">Malaria</a>	<a href="#">Research at NCBI</a>	<a href="#">VecScreen</a>
<a href="#">dbSTS</a>	<a href="#">Map Viewer</a>	<a href="#">Retroviruses</a>	<a href="#">What's New</a>
<a href="#">Education</a>	<a href="#">MGC</a>	<a href="#">SAGEmap</a>	
<a href="#">E-mail Lists</a>	<a href="#">Microbial Genomes</a>	<a href="#">Science Primer</a>	



## NCBI Web Site Search

**National Center for Biotechnology Information**  
National Library of Medicine    National Institutes of Health

PubMed   Entrez   BLAST   OMIM   Books   TaxBrowser   Structure

Search  for

**SITE MAP**  
Guide to NCBI resources  
About NCBI  
The science of our resources  
Introduction for researchers and educators

**What does NCBI do?**  
Established in 1988 as a national resource for molecular biology information, NCBI creates databases, conducts research in computational biology, develops software for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular biology.

**Hot Spots**  
Clusters of orthologous groups  
Electronic PCR  
Gene expression omnibus



# NCBI Databases and Sequence Analysis Tools



## Entrez: Search and Retrieval System

<http://www.ncbi.nlm.nih.gov/Entrez/>

The screenshot shows the NCBI Entrez search and retrieval system interface. At the top, the NCBI logo is on the left, and the text "Entrez search and retrieval system" is in the center. Below this is a navigation bar with tabs for PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, and Structure. The "Entrez" tab is selected. Below the navigation bar is a search box with a dropdown menu set to "PubMed", a "for" label, and "Go" and "Clear" buttons. On the left side, there is a sidebar with links: "About Entrez", "SITE MAP", "PubMed Help", "Entrez Help", "Entrez Tutorial NEW", "The Entrez Databases", "Batch Entrez", and "Making WWW". The main content area contains the text: "Entrez is a retrieval system for searching several linked databases. It provides access to:" followed by a list of databases with their descriptions: PubMed (biomedical literature), Nucleotide (sequence database (GenBank)), Protein (sequence database), Structure (three-dimensional macromolecular structures), Genome (complete genome assemblies), Books (BookShelf online books), Domains (conserved domains (CDD)), 3D Domains (domains from Entrez Structure), GEO (Gene Expression Omnibus), GEO Datasets (curated GEO data sets), Journals (journals in Entrez), MeSH (medical subject headings), NCBI Web Site (NCBI Web site search), OMIM (Online Mendelian Inheritance in Man), PMC (full-text digital archive of life sciences journal literature), PopSet (population study datasets), SNP (single nucleotide polymorphisms), Taxonomy (organisms in GenBank), UniGene (gene-oriented clusters of transcript sequences), and UniSTS (markers and mapping data). The NCBI logo is in the bottom right corner.

# NCBI Databases

Organisms

Genomes



mRNA



Protein

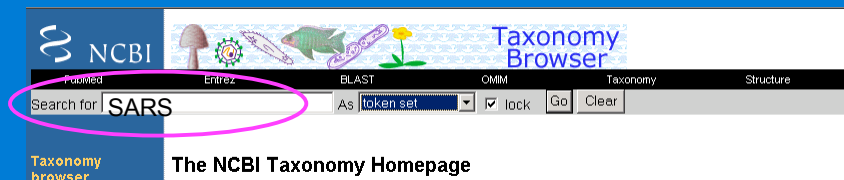


# NCBI Databases

## Organisms

<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>

incorporates phylogenetic and taxonomic knowledge from a variety of sources



# Taxonomy Browser

## SARS

- [SARS coronavirus TWH](#)
- [SARS coronavirus TWJ](#)
- [SARS coronavirus TWK](#)

### SARS coronavirus TWH

- [SARS cc](#) Taxonomy ID: 240549
- [SARS cc](#) Rank: no rank
- [SARS cc](#) Genetic code: [Translation table 1 \(Standard\)](#)
- [SARS cc](#) Lineage( full )
- [SARS cc](#) Viruses; ssRNA positive-strand viruses, no DNA stage; Nidovirales; Coronaviridae; Coronavirus;
- [SARS cc](#) SARS coronavirus
- [SARS cc](#)
- [SARS cc](#)
- [SARS cc](#) ICTV homepage

Entrez records	
Database name	Direct links
Nucleotide	<a href="#">1</a>
Protein	<a href="#">14</a>
Taxonomy	<a href="#">1</a>

### External Information Resources (NCBI LinkOut)

LinkOut	Subject	LinkOut Provider
<a href="#">NCBI SARS Resource</a>	taxonomy/phylogenetic	<a href="#">NCBI taxonomy bookmarks</a>
<a href="#">WHO</a>	taxonomy/phylogenetic	
<a href="#">CDC</a>	publishers/providers	<a href="#">National Center for Infectious Diseases</a>

Note: Groups interested in participating in the LinkOut program should visit the [LinkOut home page](#). A list of our current non-bibliographic providers can be found [here](#).



### Taxonomy browser

Archaea  
Bacteria  
Eukaryota  
Viroids  
Viruses

### Taxonomy information

Taxonomy resources

Taxonomic advisors

Genetic codes

Taxonomy Statistics

Taxonomy Name/Id Status Report

Taxonomy FTP site

FAQs

How to reference the NCBI taxonomy database

How to create links to the NCBI taxonomy

## The NCBI Taxonomy Homepage

### Taxonomy Tip of the Day

#### The pufferfish

The pufferfish genus *Fugu* has recently been renamed *Takifugu*. This is based on the recommendation of: Matsuura, K. 1990. The genus *Fugu* (Pufferfishes, Tetraodonidae) is resurrected, 1990. *Journal of Natural History Museum, London*, 19(1):1-10. Ser. A. 18(1):15-20. *Takifugu* includes the species *Takifugu rubripes*, which is a well known model organism for the vertebrate nervous system. *Takifugu rubripes* can still be retrieved by searching on *Fugu rubripes* is included in the database as a synonym of *Takifugu rubripes*.

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	<a href="#">7077719</a>	<a href="#">7077713</a>
Protein	<a href="#">191973</a>	<a href="#">191973</a>
Structure	<a href="#">4357</a>	<a href="#">4357</a>
Genome	<a href="#">25</a>	<a href="#">25</a>
Popset	<a href="#">309</a>	<a href="#">309</a>
SNP	<a href="#">4145589</a>	<a href="#">4145589</a>
3D Domains	<a href="#">15670</a>	<a href="#">15670</a>
Domains	<a href="#">11</a>	-
GEO Datasets	<a href="#">94</a>	<a href="#">94</a>
UniGene	<a href="#">108094</a>	<a href="#">108094</a>
UniSTS	<a href="#">155827</a>	<a href="#">155827</a>
PubMed Central	<a href="#">1021</a>	<a href="#">1021</a>
Taxonomy	<a href="#">2</a>	<a href="#">1</a>

These are direct links to some of the organisms commonly used in NCBI taxonomy projects:

<a href="#">Arabidopsis thaliana</a>	<a href="#">Escherichia coli</a>	<a href="#">Pneumocystis carinii</a>
<a href="#">Bos taurus</a>	<a href="#">Hepatitis C virus</a>	<a href="#">Rattus norvegicus</a>
<a href="#">Caenorhabditis elegans</a>	<a href="#">Homo sapiens</a>	<a href="#">Saccharomyces cerevisiae</a>
<a href="#">Chlamydomonas reinhardtii</a>	<a href="#">Mus musculus</a>	<a href="#">Schizosaccharomyces pombe</a>
<a href="#">Danio rerio (zebrafish)</a>	<a href="#">Mycoplasma pneumoniae</a>	<a href="#">Takifugu rubripes</a>
<a href="#">Dictyostelium discoideum</a>	<a href="#">Oryza sativa</a>	<a href="#">Xenopus laevis</a>
<a href="#">Drosophila melanogaster</a>	<a href="#">Plasmodium falciparum</a>	<a href="#">Zea mays</a>



# NCBI Databases

## Genomic DNA Sequences

Individual labs

Bulk submissions (from sequencing centers)

dbSTS, dbGSS, dbHTGS

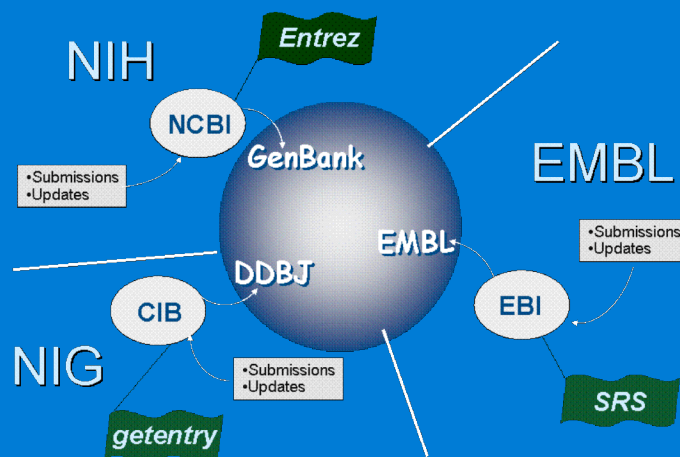
Partial/complete genomes

dbSNP



## International Nucleotide Sequence Database Collaboration

<http://www.ncbi.nlm.nih.gov/Genbank/index.html>



# Complete/Partial Genomes

<http://www.ncbi.nlm.nih.gov/Genomes/index.html>

**Genomic-scale science**

Genomics is a new and fascinating area of biology, enabled through the large-scale DNA sequencing efforts of many public and private organizations, including the [Human Genome Project](#). Genomics takes an holistic approach to molecular biology and evolution by studying the complete genome and its protein expression patterns.

**Human Genome**

Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

**The SNP Database**

Single nucleotide polymorphisms (SNPs) are the most common genetic variations and occur once every 100 to 300 bases. It is expected that SNPs will accelerate the identification of

**Organism-specific resources:**

- ▶ Fruit fly
- ▶ Human
- ▶ Malaria parasite
- ▶ Microbial Genomes
- ▶ Mouse
- ▶ Plant Genomes Central
- ▶ Rat
- ▶ Retroviruses
- ▶ Zebrafish

# Entrez Genomes

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>

[http://www.ncbi.nlm.nih.gov/genomes/static/eub\\_g.html](http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html)

**Bacteria Complete Genomes** / List 132

<a href="#">Agrobacterium tumefaciens str. C58 (Cereon)</a>	<a href="#">NC_003062</a>	2841581 bp	Oct 3 2001
<a href="#">Agrobacterium tumefaciens str. C58 (U. Washington)</a>	<a href="#">NC_003063</a>	2074782 bp	Oct 3 2001
<a href="#">Agrobacterium tumefaciens str. C58 (U. Washington)</a>	<a href="#">NC_002304</a>	2841490 bp	Dec 14 2001
<a href="#">Agrobacterium tumefaciens str. C58 (U. Washington)</a>	<a href="#">NC_003303</a>	2075560 bp	Dec 14 2001
<a href="#">Aquifex aeolicus VF5</a>	<a href="#">NC_000918</a>	1551335 bp	Sep 7 2001
<a href="#">Bacillus anthracis str. A2012</a>	<a href="#">NC_003993</a>	5093554 bp	Jun 13 2002
<a href="#">Bacillus anthracis str. Ames</a>	<a href="#">NC_003997</a>	5227293 bp	Apr 30 2003
<a href="#">Bacillus cereus ATCC 14579</a>	<a href="#">NC_004722</a>	5411809 bp	Apr 17 2003
<a href="#">Bacillus halodurans</a>	<a href="#">NC_002570</a>	4202353 bp	Sep 10 2001
<a href="#">Bacillus subtilis subsp. subtilis str. 168</a>	<a href="#">NC_000964</a>	4214814 bp	Nov 20 1997
<a href="#">Bacteroides thetaiotaomicron VPI-5482</a>	<a href="#">NC_004663</a>	6260361 bp	Mar 28 2003
<a href="#">Bifidobacterium longum NCC2705</a>	<a href="#">NC_004307</a>	2256646 bp	Sep 26 2000



http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=181661

**Agrobacterium tumefaciens str. C58 (Cereon)**

Taxonomy ID: 181661  
 Rank: no rank  
 Genetic code: [Translation table 11](#)

Other names:  
**Rhizobium radiobacter str. C58 (Cereon)**[synonym]

Lineage(full)  
[cellular organisms](#); [Bacteria](#); [Proteobacteria](#); [Alphaproteobacteria](#); [Rhizobiales](#); [Rhizobiaceae](#); [Rhizobium/Agrobacterium group](#); [Agrobacterium](#); [Agrobacterium tumefaciens](#); [Agrobacterium tumefaciens str. C58](#)

Database name	Direct links
Nucleotide	515
Protein	10600
Genome	4
Taxonomy	1

Comments and References:  
[Goodner B et al. \(2001\)](#)  
 Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Quorllo, B., Goldman, B.S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Hourmel, K., Gordon, J., Vaudin, M., Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C., and Slater, S. "Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58." Science (2001) 294:2323-2328.

**Complete genome:**  
[\[ circular \]](#) [\[ linear \]](#) [\[ plasmid AT \]](#) [\[ plasmid T1 \]](#)

[Agrobacterium tumefaciens strain C58 circular chromosome, complete sequence](#) [Microbial genomes](#)

Accession: **NC\_003062**  
 Total Bases: 2841581 bp  
 Completed: Oct 2, 2001.

Feature table:  
[Protein coding genes](#)  
[Structural RNAs](#)

**BLAST protein homologs:**  
[COGs](#) (Clusters of Orthologous Groups)  
[3D Structure](#) (Sequences with known structure)  
[TaxMap](#) (Sequences grouped by superkingdom)  
[TaxPlot](#) (3-way genome comparison)  
[CDD](#) (Conserved Domain Database)

Contributor: [Cereon Genomics](#)  
 Download chromosome sequence data from [NCBI FTP site](#)

[BLAST your query sequence against the genome](#)

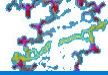
**Agrobacterium tumefaciens strain C58 circular chromosome, complete sequence**

Download from NCBI FTP site  
 RefSeq version NC\_003062 [\[Table\]](#), [\[FASTA protein\]](#), [\[FASTA nucleotide\]](#)  
 GenBank version AE007869 [\[Table\]](#), [\[FASTA protein\]](#), [\[FASTA nucleotide\]](#)

◆ - GenBank record including protein ◆ - DNA region in flatfile format ◆ - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	COG	Synonym	Product
◆ ◆ ◆	1..822	+	274	15887360		COG1806	AGR_C_5142	AGR_C_5142p
◆ ◆ ◆	797..1453	+	219	15887361		COG0424	AGR_C_2	AGR_C_2p
◆ ◆ ◆	1437..2306	+	290	15887362		COG0169	AGR_C_3	AGR_C_3p
◆ ◆ ◆	2303..2887	+	195	15887363		COG0237	AGR_C_5	AGR_C_5p

### Single Nucleotide Polymorphism



<http://www.ncbi.nlm.nih.gov/SNP/>

A central repository for

- single base nucleotide substitutions
- short deletion and insertion polymorphisms
- microsatellite repeats

Each entry includes

- Variation information
- Surrounding sequence
- Occurrence frequency
- Assay conditions



## NCBI Databases

Organisms

Genomes



mRNA



Protein



# NCBI Databases

## Expressed Sequences

mRNA

dbEST

UniGene

Gene Expression Omnibus



UniGene

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>

- an experimental system
- automatically partitioning expressed sequences
- non-redundant set of gene-oriented clusters

### Each cluster contains

- sequences that represent a unique gene
- tissue types
- map location
- selected model organism protein similarities





<http://www.ncbi.nlm.nih.gov/geo/>

- first fully public high-throughput gene expression data repository
- curated, online resource for gene expression data browsing, query and retrieval



## NCBI Databases

Organisms

Genomes



mRNA



Protein



# NCBI Databases

## Protein

- Conceptual translations of GenBank and RefSeq records
- SwissProt, PIR, PRF, PDB
- Conserved domains (CDD)



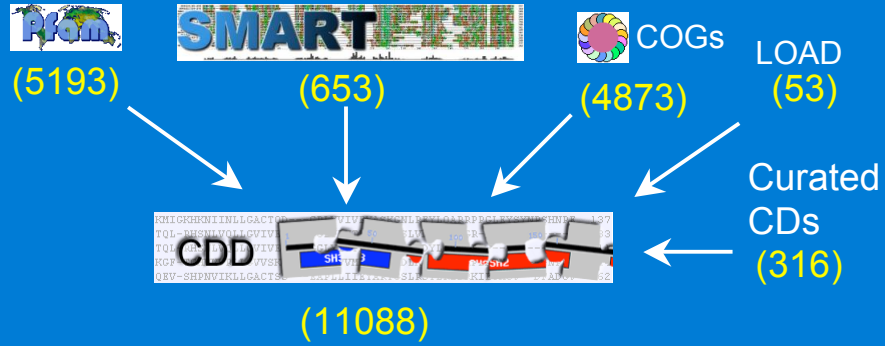
<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

## Conserved Domain

- recurring unit in molecular evolution, whose extents can be determined by sequence and structure analysis
- performs a particular function
- represented as a multiple local sequence alignment of proteins containing the domain



## Conserved Domain Database



- A position-specific scoring matrix (PSSM) is calculated
- CD-Search can be used to search against the PSSMs
- Manual curation of CDs has begun



## NCBI Databases

### Structure

- 3D Structure (MMDB)
- 3D Domains



## Molecular Modeling DataBase (MMDB)

<http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml>

- obtained from the Protein Data Bank (PDB)
- experimentally determined 3D structures
- can be viewed using Cn3D
  
- sequences also available in the Entrez protein database
- useful for finding homologs amongst known structures for a protein sequence in Entrez



## NCBI Databases

Organisms

Genomes



mRNA



Protein

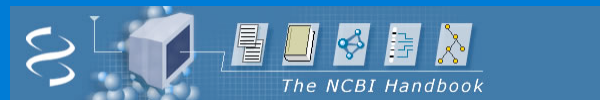


## Entrez: Search and Retrieval System for NCBI Databases

Organism	Phylogenetic and taxonomic information
Nucleotide	GenBank, EMBL, DDBJ, RefSeq, PDB
PopSet	Population study datasets
Genome	Complete genomes
SNP	Single nucleotide polymorphism
EST	Expressed sequence tags
UniGene	Clusters of expressed sequences
GEO	Microarray datasets
Protein	Translations of GenBank & RefSeq records, SWISS-PROT, PIR, PRF, PDB
Domains	CDD: conserved domain database
Structure	MMDB: experimental 3D structures
Pubmed	Biomedical literature
PubMed Central	Free online journals
Books	Free online textbooks



## Online Books





## NCBI Databases

Primary	Derived
Archival/repository	Curated
Redundant	Non-redundant
Submitter owner	NCBI owner
Sequenced	Combined/edited
Ex: GenBank	Ex: RefSeq



<http://www.ncbi.nlm.nih.gov/RefSeq/>

- best, comprehensive, non-redundant set of sequences
- for genomic DNA, transcript (RNA), and protein
- for major research organisms
  - 2005 organisms
- based on GenBank derived sequences
- ongoing curation by NCBI staff and collaborators, with review status indicated on each record
- updates to reflect current knowledge of sequence data and biology





### Partial Accession Number List

NM_123456	mRNA	
NP_123456	Protein	
NR_123456	RNA	Non-coding transcripts
NG_123456	Genomic	Incomplete genomic region
NT_123456	Genomic	BAC sequence assemblies
NW_123456	Genomic	WGS sequence assemblies
NC_123456	Genomic	Complete genomic molecules
XM_123456	mRNA	Genome Annotation
XR_123456	RNA	Genome Annotation
XP_123456	Protein	Genome Annotation



## NCBI Databases and Sequence Analysis Tools



## An Array of Sequence Analysis Tools



The Basic Local Alignment Search Tool

Search for similar protein and nucleotide seq

<http://www.ncbi.nlm.nih.gov/BLAST/>



VAST-Search – a structure-structure similarity search service

<http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>

CD-Search: Conserved Domain search tool

<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

CDART: Conserved Domain Architecture Retrieval Tool

<http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps>



## An Array of Sequence Analysis Tools



LocusLink – a single query interface for curated information about genetic loci

<http://www.ncbi.nlm.nih.gov/LocusLink/>



MapViewer – to view the complete genome assembly and annotation

<http://www.ncbi.nlm.nih.gov/mapview/>

Model Maker To generate an alternatively spliced product



Spidey – to determine exon-intron structure by aligning genomic and mRNA sequences

<http://www.ncbi.nlm.nih.gov/spidey>



VecScreen – to identify vector contamination

<http://www.ncbi.nlm.nih.gov/VecScreen/>




# BLAST Programs

blastn                    nucleotide X nucleotide  
 blastp                    protein X protein

6 frame translated nucleotide searches  
 blastx                    nucleotide X protein  
 tblastx                   nucleotide X nucleotide  
 tblastn                   protein X nucleotide


























NCBI  NCBI Map Viewer

Genome   Taxonomy   Entrez   BLAST   Help

Search  for

**New!** - searching for map objects in any user-determined subset of all plant genomes presented by NCBI. A map object includes, but is not limited to, a locus, probe name, GenBank accession, gene or name of BAC clone. Select "All plants" from the search menu at the top of the page or click (S) to the right of the Plants node.

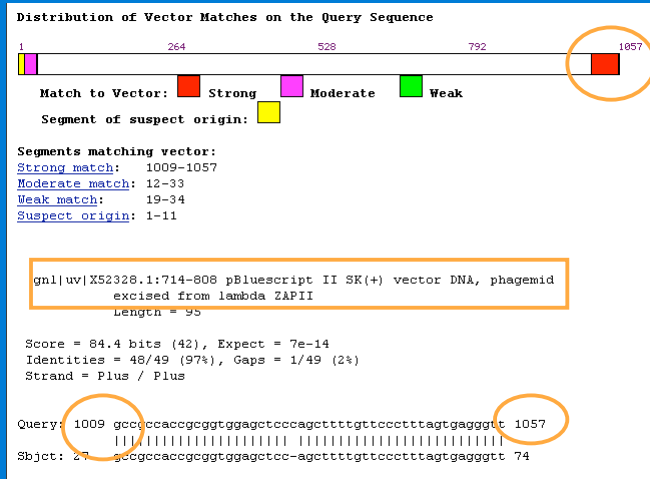
Click the  to BLAST  
 Click the  to search the group  
 Click on the binomial to view the genome overview

- Other Vertebrates**
  -  *Danio rerio* (zebrafish)
- Mammals**
  -  *Homo sapiens* (human)
  -  *Mus musculus* (mouse)
  -  *Rattus norvegicus* (rat)
- Invertebrates**
  -  *Anopheles gambiae* (mosquito)
  -  *Caenorhabditis elegans* (nematode)
  -  *Drosophila melanogaster* (fruit fly)
- Fungi**
  -  *Saccharomyces cerevisiae* (baker's yeast)
  -  *Schizosaccharomyces pombe* (fission yeast)
  -  *Neurospora crassa*
- Protozoa**
  -  *Plasmodium falciparum*
- Plants**  
  -  *Arabidopsis thaliana* (thale cress)
  -  *Avena sativa* (oat)
  -  *Hordeum vulgare* (barley)
  -  *Oryza sativa* (rice)
  -  *Triticum aestivum* (wheat)
  -  *Zea mays* (corn)
  -  *Lycopersicon esculentum* (tomato)
  -  *Glycine max* (soybean)





# VecScreen



# Spidey

**Genomic sequence (FASTA or GI/Accession):**

Upload file:

From:  To:

**mRNA sequence(s) (One or more FASTA or GI/Accession):**

Upload file:

divergent sequences  
 Use large intron sizes

Minimum mRNA-genomic identity  %  
 Minimum length of mRNA covered  %

**Genomic sequence is:**

- Vertebrate
- Drosophila
- C. elegans
- Plant

**Output options:**

- Text/summary
- Summary only
- ASN.1
- Print multiple alignment



# Spidey



Genomic: [gi2282011|gb|AC002390.1|AC002390](#) Human DNA from overlapping chromosome 19-specific cosmids R30072 and R28588, genomic sequence

mRNA: [gi21618360|ref|NM\\_014164.3](#) Homo sapiens FXYP domain containing ion transport regulator 5 (FXYP5), mRNA

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence  
mRNA coverage: 100%  
Overall percent identity: 100.0%

516 |-----| 15730

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
<a href="#">Exon 1</a>	516-657	1-142	142	100.0%	0	0	d	
<a href="#">Exon 2</a>	1399-1459	143-203	61	100.0%	0	0	d	a
<a href="#">Exon 3</a>	3269-3349	204-284	81	100.0%	0	0	d	a
<a href="#">Exon 4</a>	4192-4248	285-341	57	100.0%	0	0	d	a
<a href="#">Exon 5</a>	6557-6649	342-434	93	100.0%	0	0	d	a
<a href="#">Exon 6</a>	10004-10093	435-524	90	100.0%	0	0	d	a



## NCBI Databases and Sequence Analysis Tools



# Entrez: Search and Retrieval System

http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is a retrieval system for searching several linked databases. It provides access to:

- [PubMed](#): biomedical literature
- [Nucleotide](#): sequence database (GenBank)
- [Protein](#): sequence database
- [Structure](#): three-dimensional macromolecular structures
- [Genome](#): complete genome assemblies
- [Books](#): BookShelf online books
- [Domains](#): conserved domains (CDD)
- [3D Domains](#): domains from Entrez Structure
- [GEO](#): Gene Expression Omnibus
- [GEO Datasets](#): curated GEO data sets
- [Journals](#): journals in Entrez
- [MeSH](#): medical subject headings
- [NCBI Web Site](#): NCBI Web site search
- [OMIM](#): Online Mendelian Inheritance in Man
- [PMC](#): full-text digital archive of life sciences journal literature
- [PopSet](#): population study datasets
- [SNP](#): single nucleotide polymorphisms
- [Taxonomy](#): organisms in GenBank
- [UniGene](#): gene-oriented clusters of transcript sequences
- [UniSTS](#): markers and mapping data

**Protein** NP\_000664. class IV alcohol dehydrogenase 7 mu or sigma subunit

**Region** 8..374

**Variation** 120

**CDS** 1..374

**Location/Qualifiers**

- 1..374
- /organism="Homo sapiens"
- /db\_xref="taxon:9606"
- /chromosome="4"
- /map="iq23-q24"
- 1..374
- /product="class IV alcohol dehydrogenase 7 mu or sigma subunit"
- /EC\_number="1.1.1.1"
- /note="Alcohol dehydrogenase-7; gastric alcohol dehydrogenase"
- 8..374
- /region\_name="Zn-dependent alcohol dehydrogenases, class III [Energy production and conversion]"
- /note="AdhC"
- /db\_xref="CDD:COG1062"
- 80
- /allele="A"
- /allele="G"
- /db\_xref="dbSNP:1573496"
- 120
- /allele="H"
- /allele="R"
- /db\_xref="dbSNP:284797"
- 1..374
- /gene="ADH7"
- /coded\_by="NM\_000673.2:101..1225"
- /db\_xref="LocusID:131"
- /db\_xref="MIM:60086"

**ORIGIN**

```

1 mgtagkvik: kaavlwseqkq pfsieeieva ppktkevrik ilatgicrtd dhvikgtwms
61 kfpvivygea tgivesigeg vtvkpgdkv ipflfgpre enacnrdpnd lcirsditgr
121 gyladgtrf tckgkpvwhh mntstfeyt vdwessvaki ddaapekvk ligogfstgy
181 gaavktgkvk pgtcwrvgf gvgysvimg cksagarii gidlnkdkfe kamavgatec
241 ispkdstkpi sevlsentgn nvgytfevig hietmidala schmyy
301 ltydpllit gtwkgyvfg gkksrdvvpk lvteflakkk didqlit
361 llnsqesirt vltf
    
```





## Searching in Entrez-Nucleotide

NCBI Nucleotide

Search [Nucleotide] for [FOXP2] Preview Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Search for Genes  
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Entrez Nucleotide  
Help | FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve

- Enter terms and click Preview to see only the number of search results.
- To combine searches use # before search number, e.g., (#2 OR #3) AND asthma.

No history available

**Add Term(s) to Query or View Index:**

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

Gene Name [FOXP2] Preview Index

Click AND OR NOT to add a term to the query box.



## Searching in Entrez-Nucleotide

NCBI Nucleotide

Search [Nucleotide] for [foxp2[Gene Name]] Go Clear

Limits Preview/Index History Clipboard Details

Display [Summary] Show: 20 Send to: Text Page 1 of 4 Next

<input type="checkbox"/> 1: <a href="#">BC017397</a>	Homo sapiens forkhead box P2 (FOXP2), mRNA (cDNA clone IMAGE:4285527), complete cds [390020]	Links
<input type="checkbox"/> 2: <a href="#">NT_019482</a>	Homo sapiens forkhead box P2 (FOXP2) genomic contig [Is7_8090[29798863]]	Links
<input type="checkbox"/> 3: <a href="#">NT_039340</a>	Mus musculus chromosome 6 genomic contig, strain C57BL/6J [gi 28522947 ref NT_039340.1 Mm6_39380_30[28522947]]	Links
<input type="checkbox"/> 4: <a href="#">NM_148900</a>	Homo sapiens forkhead box P2 (FOXP2), transcript variant 4, mRNA [gi 22538414 ref NM_148900.1 [22538414]]	Links
<input type="checkbox"/> 5: <a href="#">NM_014491</a>	Homo sapiens forkhead box P2 (FOXP2), transcript variant 1, mRNA [gi 17017962 ref NM_014491.1 [17017962]]	Links

Links

- Master
- Related Sequences
- Map Viewer
- OMIM
- Protein
- PubMed
- SNP
- Taxonomy
- UniGene
- LinkOut



## Linking to Entrez-Protein

The screenshot shows the NCBI Entrez Protein search interface. The search bar contains 'NP\_055306' and the search type is set to 'Protein'. The search results show a single entry for 'NP\_055306' with a description: 'forkhead box P2 isoform I, trinucleotide repeat containing 10; forkhead/winged-helix transcription factor; speech and language disorder 1; CAG repeat protein 44 [Homo sapiens]'. A dropdown menu is open over the 'Links' section, showing options: 'Related Sequences', 'Domain Relatives', 'Map Viewer', 'Nucleotide', 'OMIM', 'PubMed', 'SNP', 'Taxonomy', and 'LinkOut'. The 'LinkOut' option is highlighted with a pink arrow.

## Outline

About NCBI

NCBI Databases and Tools

Example

EST from a hemochromatosis patient

Identify the gene

Download sequence

Known SNPs

Are they disease related

Obtain information about the gene

Conserved domain and 3D structure template

## BLAST an EST from a hemochromatosis patient library against the human genome

NCBI Genomic Biology Human Genome Guide Human Sequence

Search  for  Go

**BLAST the Human genome**

Compare your query sequence to the working draft sequence of the human genome or its mRNA and protein products.

Database:  Program:

use MegaBLAST

Enter an accession, gi, or a sequence in FASTA format:

```
TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTTTCAGTGACCACCTACGGTGTG
GGGCCCTTGAAC TACTACCCCAAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA
TGGATGCCAAGGAGTTGGAACCTAAGACGTATTGCCCAATGGGGATGGGACCTACCAGG
GCTGGATAACCTTTGGCTGTACCCCTGGGGAAAGAGCAGAGATATACGTACCAGGTGGAGC
ACCCAGGCC TGGATCAGCCCTCATTTGATCTGGG
```



## BLAST an EST from a hemochromatosis patient library against the human genome

Show positions of the BLAST hits in the human genome using the Entrez Genomes MapViewer

Query= (276 letters)

**Distribution of 1 Blast Hits on the Query Sequence**

Mouse-over to show define and scores. Click to show alignments

Color Key for Alignment Scores

<40	40-50	50-80	80-200	>200
-----	-------	-------	--------	------

Sequences producing significant alignments:

Accession	Score	E Value
U07592.13 Hs6_7749 Homo sapiens chromosome 6 genomic	525	e-147





## Download the HFE Gene Sequence

Homo sapiens Genome (build 33)  
 Region to retrieve (in chromosome coordinates):

Chromosome:  Strand:

from:  adjust by:

to:  adjust by:

Sequence Format:

---

This chromosome region corresponds to the contig region(s):

Contig	start	stop	strand
NT_007592.13	16945699	16955310	+

[+ Display](#) [Save to Disk](#) [View Evidence](#) [ModelMaker](#)



## SNPs in the HFE gene

Contig position	dbSNP rs#	cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid	
16946107	rs2858993		N.D.				intron					- Region: intron
16946540	rs807206		0.147				intron					- snp: coding
16946658	rs3798374		N.D.				intron					- snp: synonymous change
16947141	rs2784718		N.D.				intron					- snp: nonsynonymous change
16947587	rs807207		N.D.				intron					- snp: nonsynonymous change
16948644	rs2858994		N.D.				intron					- snp: untranslated region
16949254	rs2227837		0.486				intron					- snp: intron
16949347	rs2242956		N.D.		Yes		contig reference	T	Met [M]	2	35	- snp: splice-site
			N.D.		Yes		nonsynonymous change	C	Thr [T]	2	35	- snp: coding, synonymy unknown
16949430	rs1789945		0.158		Yes	Yes	contig reference	C	His [H]	1	63	
			0.158		Yes	Yes	nonsynonymous change	G	Asp [D]	1	63	
16949436	rs1800738		N.D.		Yes	Yes	contig reference	A	Ser [S]	1	85	
			N.D.		Yes	Yes	nonsynonymous change	T	Cys [C]	1	85	
18951197	rs4886951		0.012		Yes	Yes	contig reference	C	Thr [T]	2	217	
			0.012		Yes	Yes	nonsynonymous change	T	Ile [I]	2	217	
18951392	rs1800562		0.068		Yes	Yes	contig reference	G	Cys [C]	2	282	
			0.068		Yes	Yes	nonsynonymous change	A	Tyr [Y]	2	282	



## SNPs in the HFE gene

16951392	contig reference	G	Cys [C]	2	282
	nonsynonymous change	A	Tyr [Y]	2	282

```
>ref|NT_007592.13|Hs6.7749 Homo sapiens chromosome 6 genomic contig
Length = 48884767

Score = 525 hits (273), Expect = e-147
Identities = 275/276 (99%)
Strand = Plus / Plus

Query: 1      tgccctcttggggaaggtgacacatcatgtgacctcttcagtgaccactctacgggtgc 60
Sbjct: 16951164 tgccctcttggggaaggtgacacatcatgtgacctcttcagtgaccactctacgggtgc 16951223

Query: 61      gggccttgaactactacccccagaacatcaccatgaagtggtgctgaaggataagcagccaa 120
Sbjct: 16951224 gggccttgaactactacccccagaacatcaccatgaagtggtgctgaaggataagcagccaa 16951283

Query: 121     tggatgccaaggagttoaacctaaagacgtattgcccattgggatgggacctaccagg 180
Sbjct: 16951284 tggatgccaaggagttoaacctaaagacgtattgcccattgggatgggacctaccagg 16951343

Query: 181     gctggataaccttggctgtacccccctggggaagagcagagatacacgtaccagctgggagc 240
Sbjct: 16951344 gctggataaccttggctgtacccccctggggaagagcagagatacacgtaccagctgggagc 16951403
```

16951392 G:A



NCBI OMIM Online Mendelian Inheritance in Man Johns Hopkins University

Search OMIM for [Go] [Clear]

Limits Preview/Index History Clipboard De

**0001 HEMOCHROMATOSIS [HFE, CYS282TYR]**

\*235200 HEMOCHROMATOSIS; HFE

ALLELIC VARIANTS (selected examples)

- 0001 HEMOCHROMATOSIS [HFE, CYS282TYR]
- 0002 HEMOCHROMATOSIS [HFE, HIS65ASP]
- 0003 HEMOCHROMATOSIS [HFE, SER65CYS]
- 0004 HFE INTRONIC POLYMORPHISM [HFE, 5569G-A]
- 0005 HFE POLYMORPHISM [HFE, VAL53MET]
- 0006 HFE POLYMORPHISM [HFE, VAL59MET]
- 0007 PORPHYRIA VARIEGATA [HFE, GLN127HIS]
- 0008 HEMOCHROMATOSIS [HFE, ARG330MET]
- 0009 HEMOCHROMATOSIS [HFE, ILE105THR]
- 0010 HEMOCHROMATOSIS [HFE, GLY93ARG]

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>



# Outline

About NCBI

NCBI Databases and Tools

Example

EST from a hemochromatosis patient

Identify the gene (BLAST)

Download sequence (MapView)

Known SNPs (dbSNP)

Disease related (OMIM)

Obtain information about the gene

Conserved domain and 3D structure template

Why the mutant has an altered function



## Human HFE gene LocusLink Report

LocusLink Home

HFE Index: Top of Page

Nomenclature

Overview

Function

Relationships

Map

RefSeq

Related Seqs

Links

LocusLink: Collaborators

Download

FAQ

Help

Statistics

RefSeq: About

Click to Display mRNA-Genomic Alignments (spanning 9611 bps)

PUB OMIM ACBVIEW UNIGENE MAP VAR HOMOL GDB

HGMD e! UCSC

*Homo sapiens* Official Gene

Symbol and Name (HGNC)

**HFE: hemochromatosis**

LocusID: 3077

Overview ?

**RefSeq Summary:** The protein encoded by this gene is a membrane protein that is similar to MHC class I-type proteins and associates with beta2-microglobulin (beta2M). It is thought that this protein functions to regulate iron absorption by regulating the interaction of the transferrin receptor with transferrin. The iron storage disorder, hereditary haemochromatosis, is a recessive genetic disorder that results from defects in this gene. At least eleven alternatively spliced variants have been described for this gene. Additional variants have been found but their full length nature has not been determined.

**Locus Type:** gene with protein product, function known or inferred

**Product:** hemochromatosis protein isoform 1



# Human HFE gene LocusLink Report

[Home](#)  
[HFE Index:](#)  
[Top of Page](#)  
[Nomenclature](#)  
[Overview](#)  
[Function](#)  
[Relationships](#)  
[Map](#)  
[RefSeq](#)  
[Related Seqs](#)  
[Links](#)  
**LocusLink:**  
[Collaborators](#)  
[Download](#)  
[FAQ](#)  
[Help](#)  
[Statistics](#)  
[RefSeq](#)

[Function](#) [Submit GeneRIF](#) [\(All Pubs\)](#) [?](#)  
**Phenotype:**

- [Hemochromatosis](#)
- [Porphyria variegata](#)

**GeneRIF: Gene References into Function:**  
[11903354](#) • A previously undescribed nonsense mutation of the HFE gene  
[11857056](#) • Association between MHC class I gene HFE polymorphisms and longevity  
[11903355](#) • Distribution of HFE C282Y and H63D mutations in the Balearic Islands (NE Spain).  
[12148086](#) • polymorphism and its relation to type 2 diabetes mellitus in the Czech population  
[12059121](#) • Individuals with mutations in the HFE gene show very few hemochromatosis-related symptoms.



# Human HFE gene LocusLink Report

[LocusLink](#)  
[Home](#)  
[HFE Index:](#)  
[Top of Page](#)  
[Nomenclature](#)  
[Overview](#)  
[Function](#)  
[Relationships](#)  
[Map](#)  
[RefSeq](#)  
[Related Seqs](#)  
[Links](#)  
**LocusLink:**  
[Collaborators](#)  
[Download](#)  
[FAQ](#)  
[Help](#)  
[Statistics](#)

**Gene Ontology™:**

Term	Evidence	Source	Pub
<a href="#">iron ion homeostasis</a>	P	Proteome	
<a href="#">receptor mediated endocytosis</a>	E	Proteome	
<a href="#">iron ion transport</a>	E	Proteome	
<a href="#">protein complex assembly</a>	E	Proteome	
<a href="#">integral to plasma membrane</a>	E	Proteome	
<a href="#">cytoplasm</a>	E	Proteome	
<a href="#">plasma membrane</a>	E	Proteome	

**Relationships** [?](#)  
**Mouse Homology Maps:**  
 NCBI vs. MGD 13 15.00 cM [Hfe](#) [Hs](#) [Mm](#)

**Map Information** [?](#)  
**Chromosome:** 6 [mv](#)  
**Cytogenetic:** 6p21.3 RefSeq  
**Markers:** Chr. 6 [STS](#) [U60319](#) [mv](#)





# Human HFE gene LocusLink Report

[LocusLink Home](#)

**HFE Index:**  
[Top of Page](#)  
[Nomenclature](#)

[Overview](#)  
[Function](#)  
[Relationships](#)

[Map](#)  
[RefSeq](#)  
[Related Seqs](#)

[Links](#)

**LocusLink:**  
[Collaborators](#)  
[Download](#)  
[FAQ](#)  
[Help](#)  
[Statistics](#)

**RefSeq:**  
[About](#)

**Sequences (RefSeq)**

**Category:** **REVIEWED**

1. mRNA: [NM\\_000410](#)  
**Protein:** [NP\\_000401](#) hemochromatosis protein **BL** isoform 1 precursor

**Domains:** [Immunoglobulin C-Type](#) score: 152  
[Class I Histocompatibility](#) score: 314  
[antigen, domains alpha 1 and 2](#)

**Transcript Variant:** This variant (1) encodes the longest isoform.

**GenBank:** [U60319](#)

**Source:**

2. mRNA: [NM\\_139002](#)  
**Protein:** [NP\\_620571](#) hemochromatosis protein **BL** isoform 2 precursor

**Domains:** [Class I Histocompatibility](#) score: 207  
[antigen, domains alpha 1 and 2](#)

**Transcript Variant:** This variant (2) lacks the 3' end of the coding region and a portion of the 3'UTR, as compared to variant 1. The resulting protein (isoform 2) has a unique carboxy terminus.



# Hemochromatosis Protein Isoform 1 Precalculated BLAST (BL) Report

Query: gi|4504877 hemochromatosis protein isoform 1 precursor; hereditary haemochromatosis protein [Homo sapiens]  
 Matching gi: [1469790](#), [11094816](#), [22054810](#), [1890180](#), [2020069](#), [2870111](#), [2497916](#)

Best hits   Common Tree   Taxonomy Report   3D structures   **CDD-Search**   GI list

200 BLAST hits to 23 unique species [Sort by taxonomy proximity](#)

ARCHAEA    BACTERIA    METAZOA    FUNGI    PLANTS    VIRUSES    OTHER EUKARYOT

KEEP ONLY  CUT-OFF 100     

SCORE	E	ACCESSION	GI	PROTEIN DESCRIPTION
1870	27	<a href="#">AAC51823</a>	<a href="#">1469790</a>	HLA-H
1870	27	<a href="#">AAG29572</a>	<a href="#">11094315</a>	hemochromatosis termination variant terEbb
1870	25	<a href="#">AAND9793</a>	<a href="#">22854810</a>	hereditary hemochromatosis [Pan troglodyte
1870	27	<a href="#">CAB07442</a>	<a href="#">1890180</a>	haemochromatosis candidate gene [Homo sapi
1870	27	<a href="#">AAB82083</a>	<a href="#">2088551</a>	hereditary hemochromatosis [Homo sapiens]
1870	27	<a href="#">CAA20934</a>	<a href="#">2370111</a>	HFE [Homo sapiens]
1870	27	<a href="#">g30201</a>	<a href="#">2497916</a>	Hereditary hemochromatosis protein precurs
1772	27	<a href="#">AAC62646</a>	<a href="#">2675307</a>	hemochromatosis splice variant del14E4 [Hoi
1772	27	<a href="#">NP_620571</a>	<a href="#">21040347</a>	hemochromatosis protein isoform 1 precursor



# Conserved Domains in Hemochromatosis Protein Isoform 1

RPS-BLAST 2.2.6 [Apr-09-2003]

Query= gi14504377|ref|NP\_000401.1 hemochromatosis protein isoform 1 precursor; hereditary haemochromatosis protein [Homo sapiens] (348 letters)

Database: #cdd.v1.62  
11,088 PSSMs; 2,717,223 total columns

Click on boxes for multiple alignments

Show Domain Relatives

- .. THIS CD ALIGNMENT INCLUDES 3D STRUCTURE. TO DISPLAY STRUCTURE, DO

PSSMS PRODUCING SIGNIFICANT ALIGNMENTS:

- gnlCDD16636 PFAM00129, MHC\_I, CLASS I HISTOCOMPATIBILITY ANTIGEN, DO

gnlCDD16636	pfam00129, MHC_I, Class I Histocompatibility antigen, domains ...	123	3e-29
gnlCDD5323	cd00098, IgC1, Immunoglobulin domain constant region 1 (c1) su...	70.7	2e-13
gnlCDD8985	smart00407, IgC1, Immunoglobulin C-Type;	62.6	6e-11
gnlCDD16598	pfam00047, ig, Immunoglobulin domain. Members of the immunoglo...	36.8	0.004



# Proteins Containing the Conserved Domains in Hemochromatosis Protein Isoform 1

NCBI CDART: Conserved Domain Architecture Retrieval

New Query Overview Pubmed Nucleotide Protein Structure Td

Query:

Similar domain architectures:

Sequences	Protein	Description
2 Sequences	SMART00209	THROMBOSPONDIN TYPE 1 REPEATS; TYPE 1 REPEATS IN ...
2 Sequences	PFAM00090	LEUCINE-RICH REPEATS, TYPICAL (MOST POPULATED) SU...
1898 Sequences	PFAM00129	CLASS I HISTOCOMPATIBILITY ANTIGEN, DOMAINS ALPHA...
NP_111688	PFAM01463	LEUCINE RICH REPEAT C-TERMINAL DOMAIN. LEUCINE RI...
2 Sequences	SMART00082	ANIMAL HAEM PEROXIDASE.
2 Sequences	PFAM03098	SCAVENGER RECEPTOR CYS-RICH; THE SEA UCRHIN EGG P...
2 Sequences	SMART00202	DOMAIN PRESENT IN ZO-1 AND UNC5-LIKE NETRIN RECEP...
2 Sequences	PFAM00530	RETROVIRAL GAG P10 PROTEIN. THIS FAMILY CONSISTS ...
2 Sequences	SMART00218	IMMUNOGLOBULIN DOMAIN CONSTANT REGION 1 (C1) SUBF...
2 Sequences	PFAM00791	IMMUNOGLOBULIN DOMAIN CONSTANT REGION 1 (C1) SUBF...
2 Sequences	PFAM02337	LEUCINE-RICH REPEATS (LRRs), RIBONUCLEASE
2 Sequences	CD00098	IMMUNOGLOBULIN DOMAIN CONSTANT REGION 1 (C1) SUBF...

Includes: SMART00408 SMART00406 cd00097 cd00099 CD00096 SMART00407 SMART00409 PFAM00047



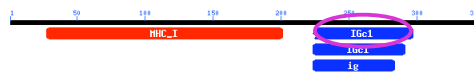
## Conserved Domains in Hemochromatosis Protein Isoform 1

RPS-BLAST 2.2.6 [Apr-09-2003]

Query= [gi145043771ref|NP\\_000401.1](#) hemochromatosis protein isoform 1 precursor; hereditary haemochromatosis protein [Homo sapiens]  
(348 letters)

Database: #cdd-v1.62  
11,088 PSSMs; 2,717-223 total columns

Click on boxes for multiple alignments



Show Domain Relatives

.. THIS CD ALIGNMENT INCLUDES 3D STRUCTURE. TO DISPLAY STRUCTURE, DC  
PSSMS PRODUCING SIGNIFICANT ALIGNMENTS:

- [gnl|CDD|16636](#) pfam09129, MHC\_I, Class I Histocompatibility antigen, domains ... [123](#) 3e-29
- [gnl|CDD|5323](#) cd00098, [IGc1](#), Immunoglobulin domain constant region 1 (c1) su... [70.7](#) 2e-13
- [gnl|CDD|8985](#) smart00407, [IGc1](#), Immunoglobulin C-Type; [62.6](#) 6e-11
- [gnl|CDD|16598](#) pfam00047, [ig](#), Immunoglobulin domain. Members of the immunoglo... [36.8](#) 0.004



## Curated Immunoglobulin Domain Constant Region 1

CD: [cd00098.1\\_IGc1](#), Query added

PSSM-Id: 5323

Source: [Smart](#)

Description: Immunoglobulin domain constant region 1 (c1) subfamily; the immunoglobulin molecule is a tetramer of two light chains and two heavy chains linked by disulfide bonds; each chain is composed of one variable domain and different numbers of constant domains; these names reflect the fact that the variability in sequences found in the variable domain is higher than that found in the constant domain; members of the c1 subfamily participate in a variety of functions and include antibodies, titin, T-cell receptors, and Major Histocompatibility Complex (MHC) class II molecules

Taxa: [Gnathostomata](#)

References: [3 Pubmed Links](#)

Related: [smart00407](#), [pfam00047](#)

Status: curated CD

Created: 25-May-2001

Aligned: 36 rows

PSSM: 99 columns

Representative: Consensus

Proteins: [Click here for CDART summary of Proteins containing cd00098](#)

View 3D Structure with [Cn3D](#) using [Virtual Bonds](#) (To display structure, [download Cn3D](#))

View Alignment as [Hypertext](#) width [60](#) color at [2.0 bits](#) feature [IG fold](#)

Subset Rows up to [10](#) sequences most similar to the query



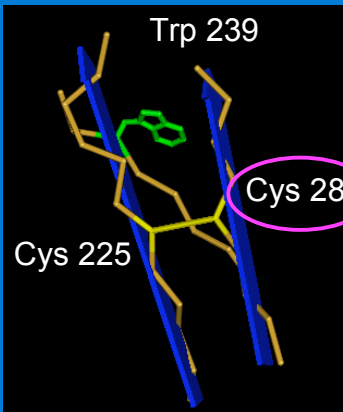
# Curated Immunoglobulin Domain Constant Region 1

**Feature 1:** IG fold  
**Evidence:** Comment: immunoglobulin fold, 2 beta-sheets forming a 'sandwich'  
 Comment: signature involves a disulfide bridge connecting the 2 beta-sheets with a Trp packing against it  
 Citation: [FMID\\_7932691](#)



## IG Fold Structure

View 3D Structure with  using  (To display structure, download [Cn3D](#))  
 View Alignment as  width  color at  feature   
 Subset Rows  sequences most similar to the query



# 3D Structural Templates for the Hemochromatosis Protein Isoform 1

Query: gi:4504377 hemochromatosis protein isoform 1 precursor; hereditary haemochromatosis protein [Homo sapiens]  
 Matching gi: 1469790, 11094815, 22264810, 1890180, 2088881, 2370111, 2497915

[Get Cn3D Now!](#)

Best hits | Common Tree | Taxonomy Report | **3D structures** | CDD-Search | GI list

114 BLAST hits to 4 unique species [Sort by taxonomy proximity](#)

ARCHAEA  BACTERIA  METAZOA  FUNGI  PLANTS  VIRUSES  OTHER EUKARYOT.

KEEP ONLY  CUT-OFF 100

348 aa	SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
	153.7	•	<a href="#">1A6ZA</a>	4659710	Chain A, Hfe (Human) Hemochromatosis Protein
	52.5	•	<a href="#">1B1IA</a>	3891929	Chain A, The Crystal Structure Of H-2dd Mhc
	50.1	•	<a href="#">1AKJA</a>	2554797	Chain A, Complex Of The Human Mhc Class I G
	49.4	•	<a href="#">1HHGA</a>	442988	Chain A, Human Class I Histocompatibility Ar
	49.0	•	<a href="#">1K5NA</a>	24987443	Chain A, Hla-B2709 Bound To Nona-Peptide M9
	48.9	•	<a href="#">1GRNA</a>	1435709	Chain A, Crystal Structure Of Human Ab Tcr C
	48.7	•	<a href="#">1KJMA</a>	27523699	Chain A, Tap-A-Associated Rat Mhc Class I M
	48.6	•	<a href="#">1G5FA</a>	6730545	Chain A, Structure Of Ab-Tcr Bound To Hla-A*



# 3D Structural Templates for the Hemochromatosis Protein Isoform 1

1A6Z - Cn3D 4.1

File View Show/Hide Style Window CDD Help

1A6Z - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

1A6Z\_A DFWT IMENHNHHSKESHTLQV I LGCEMQEDNSTEGYWKYGYDGDHLEFCPDTLDWRAAEPRAWPTKLEWERHKKI RARQRAYL  
 gi: 4504377 DFWT IMENHNHHSKESHTLQV I LGCEMQEDNSTEGYWKYGYDGDHLEFCPDTLDWRAAEPRAWPTKLEWERHKKI RARQRAYL

Human hemochromatosis protein 3D structure



# Outline

About NCBI

NCBI Databases and Tools

Example

EST from a hemochromatosis patient

Identify the gene (BLAST)

Download sequence (MapViewer)

Known SNPs (dbSNP)

Disease related (OMIM)

Obtain information about the gene (LocusLink)

Conserved domain and 3D structure template (CDD)



Information and tutorials at NCBI

BLAST Information	Nucleotide tutorial	Pubmed tutorial
Resource publications	Map Viewer exercises	Structure tutorial

Browse our science primer...

...to gain an understanding of our resources and explore our databases and tools to see what we can do for you.

**A science primer**

- Bioinformatics
- Genome Mapping
- Molecular Modeling
- SNPs
- ESTs
- Microarray Technology
- Molecular Genetics
- Pharmacogenomics
- Phylogenetics

<http://www.ncbi.nlm.nih.gov/Education/index.html>

## NCBI Core Bioinformatics Facility

- supports a network of bioinformatics specialists serving individual institutes at NIH
- trains Core Members in the use of NCBI tools
- the Core Members, in turn, support the use of NCBI's tools and databases by researchers in their institutes
- currently 16 Members from 14 institutes

Refer to the handout for the Core Member from your institute



## NCBI Training



A Field Guide to GenBank and  
NCBI Molecular Biology Resources

3 hour lecture and 2 hour hands-on

To register, go to  
<http://www.ncbi.nlm.nih.gov/Class/FieldGuide/nlm.html>



## NCBI Training

### Seven NCBI Mini-Courses

Mini-Courses 2.5 hours on specific topics  
Lecture and hands-on

Making Sense of DNA and Protein Sequences  
Unmasking Genes in Human DNA  
LocusLink Quick Start  
Structural Analysis Quick Start  
BLAST Quick Start  
MapViewer Quick Start  
GenBank and PubMed Searching



## NCBI Training

### Seven NCBI Mini-Courses

Offered at CIT and Bldg. 38A

To register  
NIH employees, go to <http://training.cit.nih.gov/>  
Non-NIH employees, send an e-mail to  
[bhagwat@ncbi.nlm.nih.gov](mailto:bhagwat@ncbi.nlm.nih.gov)

